Computational Paleography of Medieval Hebrew Scripts*

Abstract

As part of an ERC Synergy Grant entitled, *MiDRASH: Migrations of Textual and Scribal Traditions via Large-Scale Computational Analysis of Medieval Manuscripts in Hebrew Script*, we aim to develop a revolutionary approach to manuscript studies by combining traditional, digital, and computational paleographic methods to refine and perhaps rewrite our understanding of Hebrew scripts, particularly their geographical variation in scribal practices.

One of our goals is to identify, through computer sciences, clusters, and sub-clusters within different script types yet to be discovered (or those not discoverable) by paleographers. Our preliminary work focused on clustering medieval manuscripts written in Ashkenazi square script using a dataset of 206 pages extracted from 59 manuscripts. This dataset includes images of manuscripts from known origins, such as Germany and France, as well as from as yet unidentified sources. Conventional computational methods, such as the bag-of-words approach, struggle to identify the intricate features necessary for effective paleographic clustering, as the frequency of occurrence of paleographical features on a page varies even within the same script type. To address this, we had expert paleographers identify ten critical features that they use in their analyses of this script type. We trained a multi-label convolutional neural network (CNN) model to predict the presence of these features on a given page, achieving high accuracy. The ten features identified in this way are vocalization marks, left (end of line) justification, vertical stretch, strings, short descenders, fishtails, left slant, biting, nesting, and shading.

We then examined the pages globally using these features to determine if recognizable subclusters exist among them. Principal component analysis (PCA) was used to visualize the samples in 2D and identify potential clusters. By brute-forcing through all features or selected features, we found that χ^2 feature selection led to visible clusters. This feature selection process highlighted visible clusters based on the selected features (strings, left slanted, vertical stretch, and nesting), addressing the challenge faced by paleographers who can easily identify individual features on a single page but struggle to simultaneously remember and analyze these features across multiple pages to discern grouping patterns.

Looking forward, we aim to explore methods to identify discriminative features that go beyond those defined by paleographers. We assume that a script type S' possesses n distinct paleographical features that are absent in a baseline script type S (Ashkenazi square script, in our case). We train a multi-label CNN to predict the presence of all n features in images of script S', while predicting the absence of these features in images of the baseline S. Using gradient-weighted class activation mapping (Grad-CAM), we visualize the spatial locations of these n features within the images of S'. This approach enables us to identify characteristics that may not be immediately apparent to human experts, furthering our understanding of these script types.

To further enhance the representation of handwriting style features, we will incorporate another deep learning architecture. Specifically, we will train a sequence-generating recurrent neural network (RNN) on the ordered sequence of contour tip points from letter strokes. The hidden state vectors from the RNN will then be used as embedding vectors, which are expected to capture stylistic features of the handwriting.

Combining the expertise of human paleographers with advanced computational techniques, we can transcend the limitations of traditional analysis and uncover new insights into the history and development of medieval scripts. The computational analyses that we shall provide will ease the work of paleographers, enabling them to create new methodologies for the analysis of Hebrew scripts that will refine our understanding of the evolution and migration of medieval Hebrew texts. This multidisciplinary effort will result in the development of new comprehensive resources for scholars, such as a *Handbook of Medieval Hebrew Paleography*, an online *Paleographical Album*, and an encyclopedic *Vademecum of Hebrew Manuscript Studies*, thus significantly advancing the field of Hebrew paleography.

Keywords
Medieval Hebrew manuscripts, computational paleography, convolutional neural networks, image clustering, recurrent neural networks

CHR 2024: Computational Humanities Research Conference, December 4–6, 2024, Aarhus, Denmark